* We assume that $Y, X_1, \dots, X_K$ are linearly dependent and the X_i's are independent / orthogonal but when this assumption fails e.g $X_3 = X_1 + X_2$, $\det(X^T X) = 0$ so we cannot invert $X^T X$ so we cannot estimate the parameters as $\hat{\beta} = (X^T X)^{-1} X^T Y$ then we face an issue when the features are dependent or nearly dependent i.e $X_3 = X_1 + X_2 + \text{unif}(0, 1)$ (adding noise) we call this a multicollinearity problem.

* Multicollinearity also occurs when the # of features (# of cols) >> # of samples (# of rows) this is a HDLSS (high dimensional low sample size) problem which is the most difficult to fix.

* Examples are done in R (TD7).

Standard Error (SE): $\sqrt{\text{Var}(\hat{\beta})} = \sqrt{(X^T X)^{-1} MS_{Res}}$

SE = estimate of the sd of your estimate

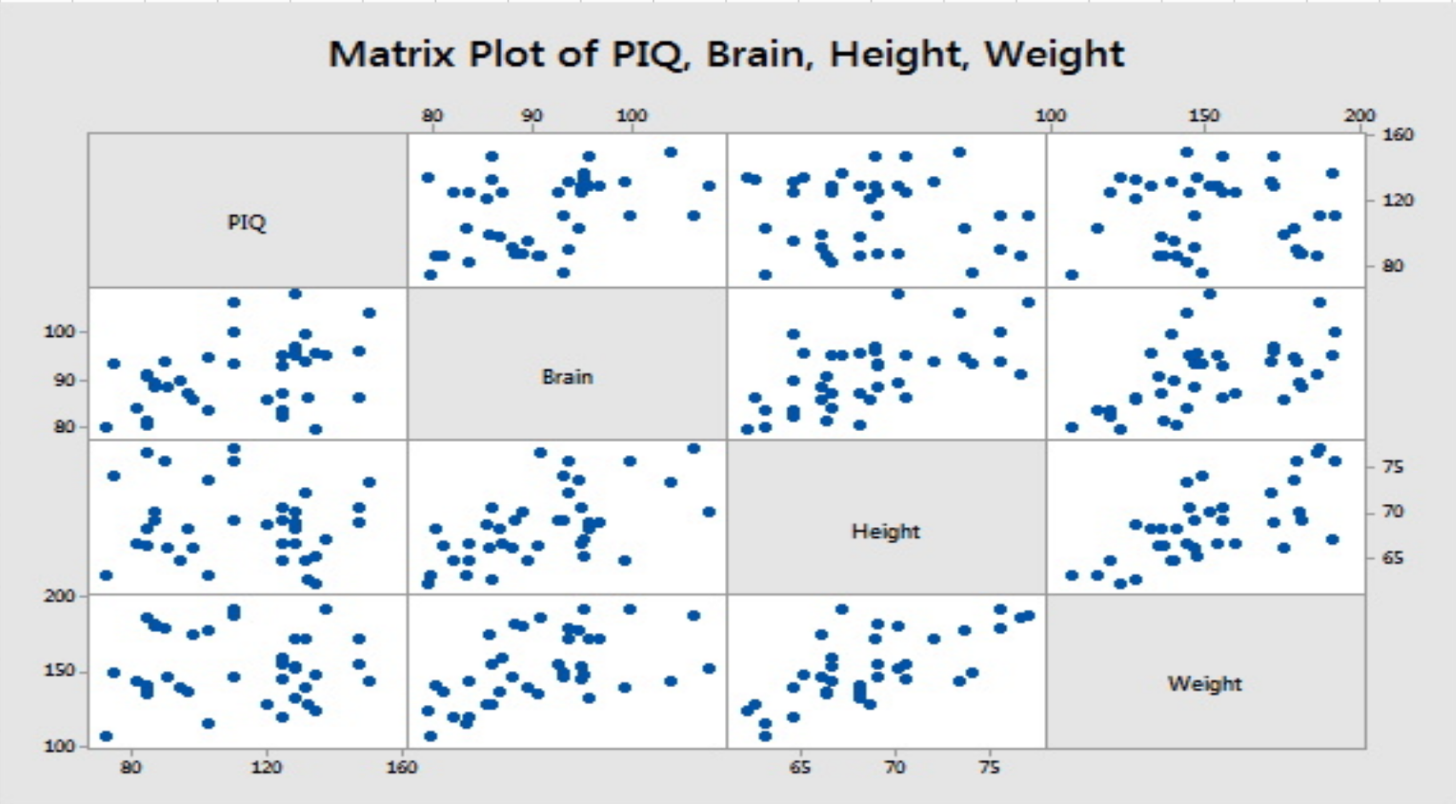
$$\sqrt{\text{Var}(\hat{\beta})} = \begin{cases} sd & , \beta \text{ is not an estimate} \\ SE & , \beta \text{ is an estimate} \end{cases}$$

* Dealing with Multicollinearity:

We use VIF (variance inflation factor), detects whether one predictor has a strong linear association with the remaining predictors with the remaining predictors.

1 : no relation > 5-10 : poor estimate

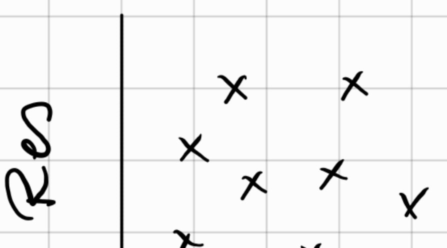
Then, we can 1) Collect additional data
2) Change estimation method
3) Deleting predictors



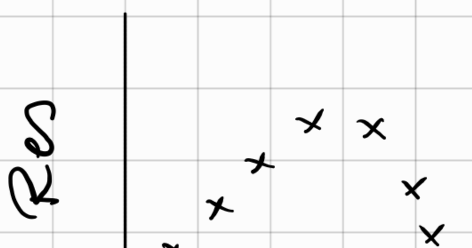
Test : $H_0 : r = 0$ vs $H_1 : r \neq 0$, $r = \rho_{xy}$

If H_0 is rejected then we have multicollinearity otherwise we don't have it.

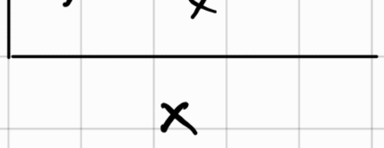
Good

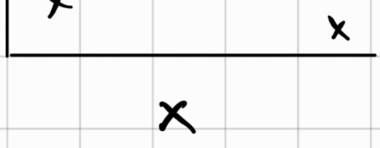


Bad

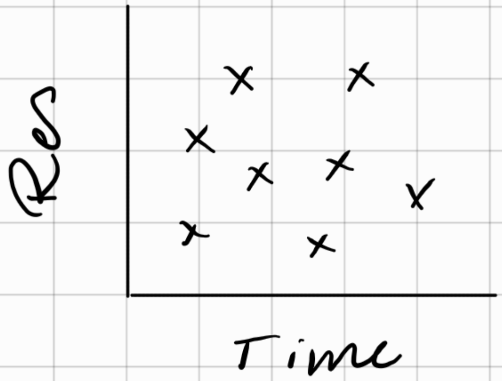


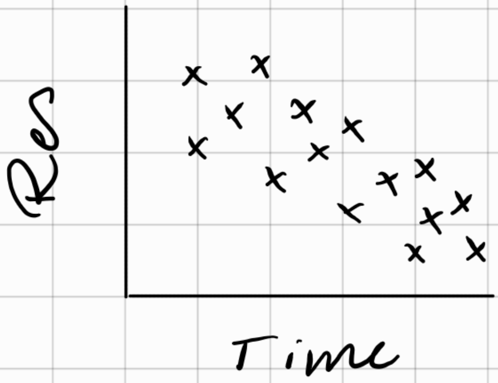
Quadratic so we can transform x to x^2 then it



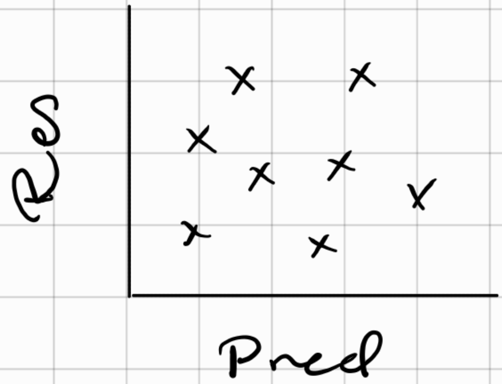


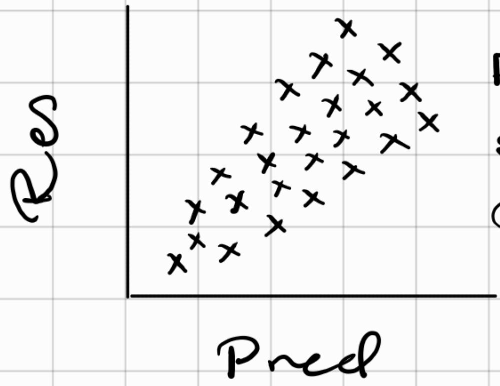
might work



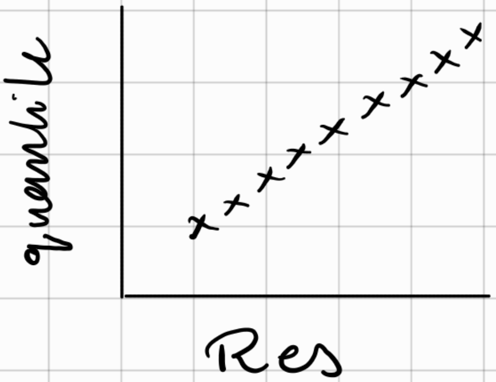


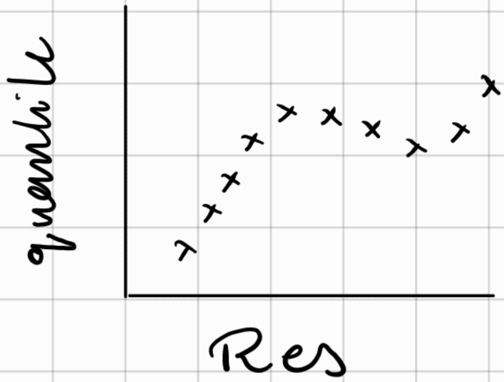
Same factors are related to time





Fan shape structure so Res ↑ as Pred ↑
log, sqrt trans.





not straight then normality assumption fails use transformation

Checking Normality

* Solving the problem of Multicollinearity :

1) Stepwise Regression : $Y, X_1, \ldots, X_k$

A) Forward Addition :

start with $Y = \beta_0 + \beta_1 X_1$ and check R^2 then consider $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and check R^2 if it increases add X_2 otherwise drop it and repeat for every X_i.

B) Backward Elimination :

Start with $Y = \beta_0 + \beta_1 X_1 + \_\_\_ + \beta_K X_K$ and check R^2 and drop X_K and check R^2, if it decreases put X_K back otherwise go on and repeat for every X_i

d) We have previous knowledge about important features so we start with $\{X_i, X_j, X_\ell\}$ and check R^2 at $Y = \beta_0 + \beta_i X_i + \beta_j X_j + \beta_\ell X_\ell$ then start by adding X_1 and check R^2 at $Y = \beta_0 + \beta_1 X_1 + \beta_i X_i + \beta_j X_j + \beta_\ell X_\ell$ if it increases keep X_1 otherwise drop it and try the rest, at the end we at least have $\{X_i, X_j, X_\ell\}$

Also, we can check adjusted R^2 and
$C_p = \frac{SSE_P}{MS_{EP}} - (n - 2p)$, SSE_P: SSE with p parameters
MS_{EP}: Mean square error

After doing this, we end up with a subset model.

* Multicollinearity exists when 2 or more covariates

Xs are strongly correlated or linearly related to each other (linearly dependent).

* Example :

X_1	X_2	X_3	Y
1	-2	4	81
2	-7	11	88
4	3	5	94
7	1	13	95

$$\lambda_1 X_1 + \lambda_2 X_2 \overset{?}{=} X_3$$

$$\begin{cases} \lambda_1 - 2\lambda_2 = 4 \\ 4\lambda_1 + 3\lambda_2 = 5 \end{cases} \Leftrightarrow \begin{cases} \lambda_1 = 4 + 2\lambda_2 \\ 16 + 8\lambda_2 + 3\lambda_2 = 5 \end{cases}$$

$$\Leftrightarrow \begin{cases} \lambda_1 = 2 \\ \lambda_2 = -1 \end{cases}$$

Hence, $X_3 = 2X_1 - X_2$ so they are correlated.

* If $(X^TX)^{-1}$ does not exist i.e X^TX is a singular matrix, we cannot get $\hat{\beta}$.

* In the previous example $\det(X^TX) = 0$ as X_3 is dependent on X_1 and X_2 (multicollinearity problem)

* Effect of Multicollinearity :

It can be shown that the diagonal elements of $(X^T X)^{-1}$ are $\frac{1}{1 - R_j^2}$; $j = 1 (1) K-1$ where R_j^2 is the coefficient of determination from the regression of X_j on the remaining $(K-2)$ regression variables

$$Var(\hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{jj} = \frac{\sigma^2}{1 - R_j^2} \xrightarrow{R_j^2 \to 1} +\infty$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \qquad X^T X = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1} x_{i2} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1} x_{i2} & \sum_{i=1}^{n} x_{i2}^2 \end{pmatrix}$$

design matrix

square and symmetric matrix

* Methods for checking collinearity :

1) Variance Inflation Factor : $VIF_i = \frac{1}{1 - R_i^2}$

$$\begin{cases} VIF \geq 5 & \Rightarrow \text{possible multicollinearity} \\ VIF \geq 10 & \Rightarrow \text{certain multicollinearity} \end{cases}$$

2) Regressing X_3 vs X_1, X_2
X_2 vs X_1, X_3
X_1 vs X_2, X_3

, if R^2 of the model is high then we have multicollinearity

Not practical for large # of features

3) Eigensystem analysis of $X^T X$: $|A - \lambda I| = 0$

- Multicollinearity detected by checking the eigenvalues of the correlation matrix $(X^T X)$.

- For the $k-1$ regressors in the model there will be $\lambda_1, \ldots, \lambda_{k-1}$

  If two or more variables are dependent then some of the eigenvalues will be really small

- Condition Number : $CN = K = \dfrac{\lambda_{max}}{\lambda_{min}}$

$$CN = \begin{cases} K < 100 & \text{no serious multicollinearity} \\ 10^2 < K < 10^3 & \text{moderate multicollinearity} \\ K > 10^3 & \text{severe multicollinearity} \end{cases}$$

* Tackling Multicollinearity :

1) Ask to get some additional data.

2) Remove the problematic Regressor

3) Combine two or more variables that are dependent into a single composite variable (collapsing variables)

4) $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$ (Ridge)

* Tips for the challenge :

1) Always check multicollinearity before modelling

2) Use Ridge regression if there is multicollinearity

3) Try Combining models :

Do Linear Regression, Ridge, Lasso, NLR, RS

Stacking Method : $\hat{Y} = \lambda_1 LR + \lambda_2 R + \lambda_3 L + \lambda_4 N + \lambda_5 RS$
and estimate λ_i's via linear regression

Take $\hat{Y}$ = mean of all models (Mixture of Expertise Models)

Take $\hat{Y}_{Linear}$ and consider it as a regressor and build another model. (Model Interm Model)

Consider interaction and split the data into training and testing data.

4) Model Selection

5) Model Adequecy $e_i \sim N(0, \sigma^2)$,